PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Modelos Geoestadísticos Utilizando Cópulas Gaussianas

Tesis para obtener el grado académico de Magíster en Estadística que presenta:

Luis Alfredo Gavidia Pantoja

Asesor:

Zaida Jesús Quiroz Cornejo

Lima, 2023

Informe de Similitud

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Modelos geoestadísticos utilizando cópulas gaussianas*, del autor Luis Alfredo Gavidia Pantoja, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 16 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 10/02/2023.
- He revisado con detalle dicho reporte y la Tesis, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 10 de febrero de 2023

Apellidos y nombres de la asesora: Quiroz Cornejo Zaida Jesús	
DNI: 43704124	Firma:
ORCID: https://orcid.org/0000-0003-3821-0815	

Dedicatoria

Dedico este trabajo a mis padres por haberme brindado su amor y cariño todos estos años.

Agradecimentos

El producto de esta tesis no es mas que la consecuencia de todas las personas que conocí en mis últimos años del pregrado, los inicios de mi vida profesional y durante el desarrollo de esta maestrìa que me alentaron seguir adelante hacia esta profesión. El agradecimiento por todo el apoyo recibido para no darme por vencido y ayudarme a retarme a mi mismo cada día. El camino no ha sido facil y sin estas personas hubiese sido imposible.

Un agradecimiendo adicional es para mi asesora de tesis por su dedicación y tiempo en este proyecto de investigación.

Resumen

La presente tesis busca aplicar una alternativa para el modelamiento de dependencia espacial de puntos georeferenciados o también conocido como datos geoestadísticos. La metodología con la que se busca abordar la autocorrelación espacial se basa en el uso de cópulas. En particular, las cópulas gaussianas brindan un marco matemático que permite definir una función de distribución conjunta acumulada a partir de la distribución marginal de la variable respuesta cuya distribución no es normal. A través de simulaciones se estudió la bondad de ajuste de los modelos geoestadísticos usando cópulas gaussianas para datos no normales. Finalmente, se aplicaron los modelos a dos bases de datos reales: i) para detectar yacimientos petrolíferos y ii) para estimar el nivel de contaminación en el aire.

Palabras-clave: Autocorrelación espacial, cópulas gaussianas, geoestadística, modelos lineales generalizados.

Abstract

This thesis applies an alternative to modelling spatial dependence of geo-referenced points also known as geostatistics data. The methodology focus on the development of spatial autocorrelation is based on using copulas. In particular, Gaussian copulas allow a mathematical framework to define a joint cumulative distribution function based on the marginal distribution of the response variable that is non gaussian. The goodness of fit of the geostatistical models using Gaussian copulas for non-normal data was studied through simulations. Finally, the models were applied to two real databases: i) to detect oil fields and ii) to estimate the level of air pollution.

Keywords: Copulas, geostatistics, generalized linear models, spatial autocorrelation.

Índice general

1. Introducción **1**

1.1. Objetivos 2

1.2. Organización del trabajo 2

2. Revisión de literatura **3**

2.1. Cópulas 3

2.1.1. Nociones Preliminares 3

2.1.2. Definición de cópula 3

2.1.3. Teorema de Slark 4

2.1.4. Función de densidad conjunta usando una cópula 5

2.1.5. Cópula gaussiana 6

2.2. Modelos Lineales Generalizados - MLG 7

2.2.1. Regresión logística 7

2.2.2. Regresión de poisson 8

2.2.3. Regresión Gamma 8

2.3. Regresión Beta 9

2.4. Geoestadística 9

2.4.1. Proceso Espacial 10

2.4.2. Proceso gaussiano 11

2.4.3. Propiedades de procesos espaciales 11

3. Modelo geoestadístico con cópula gaussiana **13**

3.1. Definición del modelo 14

3.1.1. Inferencia del Modelo 15

4. Estudio de Simulación **17**

4.1. Detalle de la generación de errores aleatorios espaciales 18

4.2. Resultados de las simulaciones 20

5. Aplicaciones **22**

5.1. Aplicación 1 22

5.2. Aplicación 2 25

6. Conclusiones **30**

6.1. Sugerencia para investigaciones futuras 31

A. Código R **32**

A.1. Simulación 32

A.1.1. Simulación Distribución Benroulli 32

A.2. Modelo Gamma 35

Bibliografía **38**

Capítulo 1

Introducción

En distintas áreas como la ecología, sociología, salud, economía, entre otras se tiene información a cerca de la ubicación de la unidad de análisis. Esta información respecto a su localización puede representarse de muchas formas. Así, Cressie (1993) sugirió por lo menos la existencia de tres tipos de datos espaciales.

En primer lugar, los datos espaciales puede ser puntos geo-referenciados. Así, sea, por ejemplo $Y(s)$ una variable aleatoria que depende, entre otras cosas, de su ubicación geográfica s de modo tal que $s \in \mathbb{R}^r$, donde s varía continuamente en D. Esto es un subespacio $\mathbb{R}^r$. En el segundo tipo de datos, se utiliza el mismo subespacio D, sin embargo ahora este queda dividido en n partes y cada una de estas será considerada como un área. Finalmente, en el tercer tipo de datos se regresa a la situación inicial de puntos; sin embargo, ahora el subespacio D será aleatorio. Así su índice responderá a patrones de ubicaciones. Para efectos de la actual investigación se utilizan los datos correspondientes al primer tipo de datos.

Una vez establecido el tipo de datos para la investigación es necesario analizar sus propiedades. Dentro de la estadística estos datos poseen características interesantes dentro de su matriz de varianzas y covarianzas. Esta no está compuesta por valores constantes debido a la existencia de la autocorrelación espacial entre observaciones. En este sentido, la autocorrelación espacial puede ser incluida en el modelo de dos formas en un modelo estadístico: i) se definen efectos aleatorios, y se asume que siguen una distribución normal multivariada (Cressie, 1993), o ii) se define una distribución marginal para la variable respuesta y se usa cópulas para definir la función de probabilidad (o densidad) conjunta (Masarotto y Varin, 2012). En cualquiera de estas propuestas, la matriz de varianzas y covarianzas permitirá definir una estructura para modelar la autocorrelación espacial.

En esta tesis se sigue la segunda propuesta. Gracias al uso de cópulas se puede modelar sin las restricciones de normalidad y linealidad que enfrenta un modelo de regresión simple. Del

mismo modo, se evita que los datos sean transformados y se puede trabajar con los mismos en su forma original. En particular, Masarotto y Varin (2012) proponen utilizar cópulas gaussianas, donde una cópula gaussiana permite definir la estructura de dependencia espacial entre varias distribuciones marginales univariadas. La cópula gaussiana permite modelar la autocorrelación entre las observaciones, por ejemplo en series de tiempo para modelar la correlación temporal en la tasa de desempleo (Gonzales, 2022). En particular, en esta tesis la cópula gaussiana permitirá incluir la autocorrelación espacial de la variable respuesta en dos ubicaciones. La estimación de los parámetros de este modelo espacial usando cópulas gaussianas se puede realizar usando inferencia clásica (Masarotto y Varin, 2017).

1.1. Objetivos

El objetivo principal de esta investigación es mostrar aplicaciones de regresiones utilizando cópulas gaussianas para datos espaciales. Así, de forma específica se busca abordar cuatro puntos.

Los objetivos específicos son:

- En primer lugar, se hace una revisión de la literatura del modelo para el análisis espacial. Como se mencionó en la sección anterior, el enfoque será en el modelo para datos georeferenciados.
- En segundo lugar, se estudiará los métodos de inferencia para el modelo espacial usando de cópulas gaussianas.
- Como tercer paso, se validará la inferencia a través de la simulación del modelo propuesto.
- Finalmente, se realizarán aplicaciones del modelo a datos con distribuciones no normales.

1.2. Organización del trabajo

El presente trabajo se organiza de la siguiente forma. En primer lugar, en la siguiente sección se realizará una breve descripción de la literatura sobre el uso de cópulas en modelos de regresión. En el siguiente capítulo se detallan los principales conceptos del uso de cópulas y geoestadística para despues explicar la inferencia del modelo así como modelos candidatos. En tercer lugar, se realiza un estudio de simulación para los modelos analizados. Finalmente, se presenta aplicaciones prácticas utilizando datos reales.

Capítulo 2

Revisión de literatura

2.1. Cópulas

2.1.1. Nociones Preliminares

Sea X una variable aleatoria (v.a.) y U una variable aleatoria con distribución uniforme continua $U(a, b)$, donde a y b son constantes. La función de distribución acumulada (fda) de X es dada por $F_X(x) = P(X < x)$ donde $F : \mathbb{R} \to [0, 1]$. Luego, $F_X(x) = u$ donde u es una valor de la v.a. $U(0, 1)$, así cualquier variable aleatoria X puede ser definida a través de la fda de X y una distribución $U(0, 1)$. Se podrían aplicar simulaciones de la variable aleatoria X aplicando la inversa de su función de distribución acumulada $F_X(x)$ y se obtenbdría:

$$x = F_X^{-1}(u).$$

Vale la pena notar que la distribución uniforme se caracteriza por poseer una propiedad lineal, es decir, $P(U < u) = u$, esto se prueba de la siguiente forma:

$$P(U < u) = P(F_X(x) < u) = P(X < F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u.$$

2.1.2. Definición de cópula

Una cópula permite analizar la dependencia entre variables aleatorias a partir de sendas distribuciones acumuladas. De acuerdo con Kurowicka y Cooke (2006) pueden especificar de forma separada los efectos de las distribuciones marginales y efectos de la dependencia.

Formalmente, una cópula n-dimensional es una función de distribución multivariada

$$C : [0, 1]^n \to [0, 1],$$

donde sus distribuciones marginales se definen por $U_i \sim U(0,1), i = 1, 2, ..., n$ (Nelsen, 2007). Dicha función cumple las siguientes propiedades:

- $C(u_1, u_2, \ldots, u_n)$ es creciente para cada componente de u_i.
- $C(1, \ldots, 1, u_i, 1, \ldots, 1)$ para todo $i = 1, \ldots, n$.
- Para todo $(u_{11}, \ldots, u_{n1})$, $(u_{12}, \ldots, u_{n2}) \in [0,1]^n$ con $u_{i1} \leq u_{i2}$, se tiene que:

$$\sum_{i1=1}^{2} ... \sum_{in=1}^{2} (-1)^{i_1+...+i_n} C(u_{1i_1}, ..., u_{ni_n}) \geq 0.$$

Las segunda condición propuesta por Nelsen (2007), es la cual sugiere que las distribuciones de U_i sean uniformes $U(0,1)$. En particular, la cópula puede ser expresada de la siguiente forma:

$$C(u_1, u_2, \ldots, u_n) = P(U_1 \leq u_1, \ldots, U_n \leq u_n),$$

la cual es la función de distribución conjunta de $U_1, U_2, \ldots, U_n$.

Para hallar la fdp de una cópula $C(u_1, ..., u_n)$ se tiene que derivar n veces respecto de $u_1, \ldots, u_n$:

$$c(u_1, ..., u_n) = \frac{d^n C(u_1, ..., u_n)}{du_1 ... du_n}.$$

De acuerdo a la definición de la cópula es posible reescribir cualquier tipo de variable X a través de una función de distribución acumulada uniforme como se explicó en la sección anterior. Esta relación nos da espacio para utilizar el teorema de Slark.

2.1.3. Teorema de Slark

Sea el vector aleatorio $(X_1, X_2, \ldots, X_n)^\top$ con una determinada función de distribución conjunta:

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

y la función de distribución acumulada (fda) marginal de X_j para $j = 1, 2, \ldots, n$ es dada por:

$$F_{X_j}(x) = F_j(x) = P(X_j \leq x).$$

Entonces existe una cópula $C : [0,1]^n \to [0,1]$ de modo tal que:

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)). \tag{2.1}$$

Es decir la función de distribución conjunta de $(X_1, X_2, \ldots, X_n)^\top$ definida previamente es una función de distribución conjunta de las fda marginales $F_j(x_j)$.

Para probar la ecuación 2.1 se define $U_j = F_j(X_j) \sim U(0,1)$, debido a que $P(U_j \leq u) = u$, entonces se puede demostrar que:

$$F(x_1, ..., x_n) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n)$$

$$F(x_1, ..., x_n) = P(F_1^{-1}(U_1) \leq x_1, F_2^{-1}(U_2) \leq x_2, ..., F_n^{-1}(U_n) \leq x_n)$$

$$F(x_1, ..., x_n) = P(U_1 \leq F_1(x_1), ..., U_n \leq F_1(x_1))$$

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)).$$

Utilizando el teorema de Slark, también se puede llegar a la conclusión que la ecuación 2.1 se puede definir como una función de la distribución uniforme con lo que, se tiene una cópula C la cual es única:

$$C(x_1, ..., x_n) = F(F_1^{-1}(u_1), ..., F_p^{-1}(u_n)). \tag{2.2}$$

2.1.4. Función de densidad conjunta usando una cópula

Asimismo, si las funciones $F(.)$ y $C(.)$ son diferenciables, entonces la función de densidad conjunta del vector aleatorio $(X_1, X_2, \ldots, X_n)^\top$:

$$f(x_1, ..., x_n) = c(F_1(x_1), .., F_n(x_n)) \prod_{i=1}^{n} f_i(x_i), \tag{2.3}$$

donde $c(.)$ es la función de densidad de la cópula y $f_i(x_i) = f_{X_i}(x_i)$.

A continuación se prueba este resultado. Como:

$$f(x_1, ..., x_n) = \frac{d^n F(x_1, ..., x_n)}{dx_1 ... dx_n} = \frac{d^n C(F(x_1, ..., x_n))}{dx_1 \ldots dx_n},$$

entonces

$$f(x_1, ..., x_n) = \frac{d^n C(F(x_1, ..., x_n))}{dF(x_1) ... dF(x_n)} \frac{dF_1(x_1)}{dx_1} \cdots \frac{dF_n(x_n)}{dx_n}.$$

Luego:

$$f(x_1, ..., x_n) = \frac{d^n C(F_1(x_1), ..., F_n(x_n))}{dF_1(x_1)...dF_n(x_n)} f_1(x_1)...f_n(x_n),$$

entonces:

$$f(x_1, ..., x_n) = c(F_1(x_1), ..., F_n(x_n)) f_1(x_1)...f_p(x_n).$$

Finalmente, la función de densidad se puede definir como:

$$f(x_1, ..., x_n) = c(F_1(x_1), .., F_n(x_n)) \prod_{i=1}^{n} f_i(x_i).$$

2.1.5. Cópula gaussiana

De acuerdo al teorema de Slark analizado en la sección anterior, una cópula gaussiana podría ser reescrita de la siguiente forma:

$$C(u_1, ..., u_n) = \Phi_\Sigma(\Phi^{-1}(u_1), .., \Phi^{-1}(u_1)), \tag{2.4}$$

en donde Φ^{-1} es la inversa de la función de distribución normal estándar y Φ_Σ es la función de distribución normal acumulada n-variada con media cero y matriz de correlación Σ. Por lo tanto la función de densidad de una cópula gaussiana se escribe de la siguiente manera:

$$c(u_1, ..., u_n) = \frac{d^n}{du_1 du_2 ... du_n} C(u_1, .., u_n)$$

$$c(u_1, ..., u_n) = \frac{d^n}{du_1 du_2 ... du_n} \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)$$

$$c(u_1, ..., u_n) = \frac{\phi_n(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))}{\phi(\Phi^{-1}(u_1))...(\Phi^{-1}(u_n))} \tag{2.5}$$

$$c(u_1, ..., u_n) = \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\Phi^{-1}(u_1), ..., \Phi^{-1}(u_1))^T (\Sigma^{-1} - I)(\Phi^-1(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_n)) \right\}, \tag{2.6}$$

donde $\phi_n(.)$ representa la función de densidad de una normal multivariada con media cero y matriz de covarianza Σ y $\phi()$ representa la fdp de una normal multivariada estándar.

Si se aplica el teorema de Slark usando esta cópula gaussiana, se puede hallar la fdp conjunta de un vector aleatorio $(X_1, X_2, \ldots, X_n)^\top$, el resultado es el siguiente:

$$f(x_1, .., x_n) = \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\Phi^{-1}(u_1), ..., \Phi^{-1}(u_1))^T (\Sigma^{-1} - I)(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_n)) \right\} \prod_{i=1}^{n} f_i(x_i).$$

2.2. Modelos Lineales Generalizados - MLG

Sea un vector aleatorio $(Y_1, Y_2, \ldots, Y_n)^\top$ donde cada Y_i es una variable aleatoria en el local s_i. Se asume que la distribución marginal de Y_i es una distribución que pertenece a la familia exponencial; es decir, $Y_i \sim FE(\mu_i, \kappa_i)$ dónde $E(Y_i) = \mu_i$ y κ_i es un parámetro de escala o dispersión. Ejemplos de las familias exponenciales son la distribución bernoulli, binomial, poisson, gaussiana, gamma, etc. Se define el MGL a través de funciones de enlace $g_1(.)$ y $g_2(.)$ que asocian a la media μ_i y κ_i a sus respectivos predictores lineales. En particular se asume:

$$g_1(\mu_i) = logit(\mu_i) = x_i'\beta_x$$

$$g_2(\kappa_i) = logit(\kappa_i) = z_i'\beta_z,$$

donde x_i es un vector de covariables de dimensión h, z_i es un vector de covariables de dimension r y β_x asi como β_z son vectores de coeficientes de regresión. A continuación a manera de ejemplo se presentan algunos modelos conocidos.

2.2.1. Regresión logística

La regresión logística se usa para modelar datos que son proporciones. En particular sea Y_i v.a. que representa el número de éxitos de n_i intentos independientes. Si π_i es la probabilidad de que ocurra éxito en cada uno de estos intentos. Luego se asume que Y_i distribuida de forma independiente como una binomial de la forma:

$$Y_i \overset{ind}{\sim} B(\pi_i, n_i).$$

Como n_i es conocido se usan variables explicativas para estimar la probabilidad de ocurrencia de cada uno de estos eventos π_i. Como π_i está definido dentro del intervalo [0,1], se puede usar la función de enlace g_1 como una función logit. Así, el modelo queda definido de la siguiente forma:

$$g_1(\pi_i) = logit(\pi_i) = x_i'\beta_x.$$

Luego la probabilidad de éxito para cada i-ésimo individuo o sujeot puede ser expresada de la siguiente manera:

$$\pi_i = \frac{\exp(x_i'\beta_x)}{1 + \exp(x_i'\beta_x)}.$$

2.2.2. Regresión de poisson

La regresión poisson tiene como variable dependiente una variable de conteo que se modela a través de la distribución de poisson:

$$Y_i \overset{ind}{\sim} Poisson(\lambda_i),$$

en donde $\lambda_i = E(Y_i) > 0$. Usando una función de enlace logarítmica,

$$g_1(\lambda_i) = \log(\lambda_i) = x_i'\beta_x.$$

Luego,

$$\lambda_i = \exp(x_i'\beta_x).$$

Así, la función de máximo de verosimilitud es de la siguiente forma:

$$L(\beta_x) = \prod_{i=1}^{n} f(y_i|x_i, \beta_x) = \prod_{i=1}^{n} \frac{e^{-\exp(x_i'\beta_x)} \exp(x_i'\beta_x)}{y_i!}.$$

2.2.3. Regresión Gamma

La regresión gamma tiene como variable dependiente una variable continua positiva que se modela a través de una distribución gamma:

$$Y_i \sim Gamma(\mu_i, \kappa),$$

en donde $\mu_i > 0$ representa el parámetro de forma mientras que $\kappa > 0$ es un parámetro de dispersión.

Ahora bien la media puede ser enlazada al predictor lineal a través de una función de enlace logarítmica de la siguiente forma:

$$g_1(\mu_i) = \log(\mu_i) = x_i'\beta_x.$$

Luego,

$$\mu_i = \exp(x_i'\beta_x).$$

2.3. Regresión Beta

La distribución beta no pertenece a la familia exponencial. Sin embargo el modelo de regresión beta se plantea de forma muy similar a un MLG. Es una regresión útil para modelar datos en el intervalo (0,1).

Sea Y v.a. que toma valores en el intervalo (0,1) tal que se asume que sigue una distribución beta de la forma:

$$Y \sim beta(p, q).$$

Con lo cual se tendría la siguiente función de densidad:

$$f(y : p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}.$$

De acuerdo a Cribari-Neto (2006) es posible reparametrizar esta distribución $Y \sim beta(\mu, \kappa)$ de la siguiente forma:

$$\mu = \frac{p}{(p+q)}$$

$$\kappa = p + q.$$

Así, en un modelo de regresión beta, $Y_i \overset{ind}{\sim} beta(\mu_i, \kappa)$, luego la fdp de Y_i queda representada de la siguiente forma:

$$f(y_i : \mu_i, \kappa) = \frac{\Gamma(\kappa)}{\Gamma(\mu_i\kappa)\Gamma((1-\mu_i)\kappa)} y_i^{\mu_i\kappa-1}(1-y_i)^{(1-\mu_i)\kappa-1}.$$

De este modo $E(Y_i) = \mu_i$ y $var(y) = \frac{\mu_i(1-\mu_i)}{1+\kappa}$. Como μ_i está definido dentro del intervalo (0,1), se puede usar la función de enlace g_1 como una función logit. Así, el modelo queda definido de la siguiente forma:

$$g_1(\mu_i) = logit(\mu_i) = \log(\mu_i/(1-\mu_i)) = x_i'\beta_x.$$

2.4. Geoestadística

La geoestadística es el área de la estadística espacial que modela datos geo-referenciados. Los datos se dicen que son geo-referenciados cuando se tiene información sobre una variable Y y el lugar de referencia donde fue recolectado. Así, se puede estudiar por ejemplo la temperatura en el local s_i que representa la latitud y la longuitud. En la práctica se recolectan los datos en un número finito de locales $i = 1, 2, ..n$; sin embargo, la variable de interés es estudiada

en un dominio continuo $D \in \mathbb{R}^n$, por ello formalmente se define un proceso espacial:

$$\{Y(s) : s \in D\}.$$

2.4.1. Proceso Espacial

Se dice que un proceso espacial es definido como $\{Y(s) : s \in D\}$, entonces $Y(s)$ es una variable aleatoria cuya función de distribución acumulada se va a representar de la siguiente forma:

$$F_{Y_{(s)}}(y) = P(Y(s) \leq y),$$

es decir una variable aletoria estudiada tome un valor menor o igual que el número real y en una definición de s. De igual modo, $(Y(s_1), Y(s_2))$ es un vector aleatorio, entonces su función de distribución acumulada conjunta es bivariada, de la forma:

$$F_{Y_{(s_1)},Y_{(s_2)}}(y_1, Y_2) = P(Y(s_1) \leq y_1, Y(s_2) \leq y_2),$$

esto define la probabilidad que una variable aleatoria tome un valor menor o igual al numero real y_1 en el local s_1 y en la misma realización menor o igual que y_2 en el local s_2. Así, esta definción se puede extender hacia distribuciones de vectores aleatorios con n componentes, para cualquier n. De ese modo, todas las familias de distribuciones incluidas están definidas en un proceso espacial. En particular, un proceso gaussiano se va a definir como una familia de distribuciones finito-dimensionales.

Ahora bien, se pueden definir momentos de una distribución espacial. Estos son las siguientes funciones donde se asume que $Y(s) = Y_s$ y las funciones dependen del local:

1. **Media**:
$$\mu_s = E(Y_s)$$

2. **Varianza**:
$$\sigma_s^2 = V(Y_s)$$

3. **Covarianza**:
$$Cov(s_i, s_j) = Cov(s_j, s_s) = Cov(Y_{s_i}, Y_{s_j})$$

4. **Correlación**:
$$\rho(s_j, s_s) = \frac{Cov(Y_{s_i}, Y_{s_j})}{\sigma_{s_i}^2 \sigma_{s_j}^2}$$

2.4.2. Proceso gaussiano

Un proceso gaussiano $\{Y(s), s \in D\}$ es definido por una familia de distribuciones finito-dimensionales gaussianas:

$$Y(s) \sim PG(\mu_s, C(h)),$$

en donde μ_s es la función de medias y $C(h)$ es la función de covarianzas, las cuales dependen de la distancia h determinado con los locales s. Así, la realización del proceso gaussiano se puede ver de la siguiente forma:

$$(Y(s_1), Y(s_2), \ldots, Y(s_n))^\top \sim N_n(0, \Sigma).$$

2.4.3. Propiedades de procesos espaciales

De acuerdo a Cressie (1993) es posible definir las propiedades siguientes en los procesos espaciales:

Estacionariedad

Existen dos tipos de estacionariedad en los procesos espaciales: fuerte y débil. En el primer caso, se dice que un proceso es estacionario, siempre y cuando una cantidad de ubicaciones $s_1, s_2 ... s_n$ y para cualquier distancia $h \in \mathbb{R}^r$, la distribución $(Y(s_1), Y(s_2), \ldots, Y(s_n))^\top$ es la misma que $(Y(s_1 + h), Y(s_2 + h), \ldots, Y(s_n + h))^\top$. En el caso de la estacionariedad débil, se cumple que

$$\mu(s) = \mu,$$

es decir la media es constante para toda ubicación s, además la

$$Cov(Y(s), Y(s + h)) = C(h),$$

para cualquier h.

Intrínsecamente estacionario

Existe un tercer tipo de estacionariedad, esta implica que se cumpla la siguiente relación:

$$E[Y(s + h) - Y(s)]^2 = Var(Y(s + h) - Y(s)) = 2\gamma(h),$$

en donde la función $2\gamma(h)$ se llama variograma y $\gamma(h)$ es llamado semivariograma.

Asímismo, existe una relación entre el variograma y la función de covarianzas C(.):

$$2\gamma(h) = Var(Y(s+h)) + Var(Y(s)) - 2Cov(Var(Y(s+h), Y(s))$$

$$2\gamma(h) = C(0) + C(0) - 2C(h)$$

$$\gamma(h) = C(0) - C(h).$$

Isotropía

Si el semivariograma $\gamma(h)$ solo depende de la distancia h, se dice que el proceso es isotrópico. Por ejemplo, el semivariograma exponencial de un proceso estacionario es intrínsecamente estacionario e isotrópico, y se define de la siguiente forma:

$$\gamma(h) = \tau^2 + \sigma^2(1 - \exp(-\phi h)),$$

donde $\tau^2 > 0$ el cual se conoce como efecto pepita, mientras que ϕ y σ^2 son parámetros de decaimiento y varianza marginal, respectivamente. La matriz de covarianza exponencial puede definirse a través de:

$$C(h) = C(0) - C(h) = \lim_{u \to \infty} \gamma(u) - \gamma(h)$$

$$C(h) = \tau^2 + \sigma^2 - [\tau^2 + \sigma^2(1 - \exp(-\phi h))]$$

$$C(h) = \sigma^2 \exp(-\phi h).$$

Capítulo 3

Modelo geoestadístico con cópula gaussiana

Como ya se ha mencionado, los modelos de regresión lineal suelen tener limitaciones. Estas pueden tener problemas de no normalidadd, no linealidad, etc. Así la literatura ha propuesto distintos modelos para corregir estas limitaciones.

Joe (2014) sugiere una regresión utilizando cópulas el cual tiene como enfoque la especificación de la cópula como modelo de regresión así como la estructura para poder definir su dependencia. Kazianka y Pilz (2010) por otro lado, señala que en geoestadística se puede utilizar el variograma para describir la dependencia espacial. Sin embargo, estos son altamente influidos cuando hay observaciones periféricas muy grandes (valores extremos). Entonces, los autores sugieren que con el uso de cópulas se pueden corregir estos problemas.

Bai et al. (2014) utiliza información espacial en conglomerados. Para ello, utiliza una metodología llamada geocópula. La cual se aplica a datos binarios, continuos y de conteo. Para la inferencia se utiliza un eficiente enfoque de verosimilitud compuesta (composite likelihood).

Pitt et al. (2006) indica que un modelo de regresión a través de cópulas permite el uso de datos cuya distribución no necesariamente es gaussiana, por ejemplo a partir de distribuciones marginales discretas y continuas. Propone utilizar un enfoque bayesiano para estimar los parámetros. Mientras Song et al. (2009) busca una aproximación conjunta para datos correlacionados. Su aproximación permite obtener una ganancia de eficiencia en la estimación de parámetros.

Finalmente, Krupskii y Genton (2018) proponen una metodología de cópulas para data espacial multivariada. La cópula propuesta se basa en el supuesto que existen algunos factores que afectan la dependencia espacial conjunta de todas las mediciones de cada variable así como al dependencia conjunta entre variables. Este modelo se parametriza en términos de una

función de covarianza cruzada así como factores aditivos que permiten analizar la dependencia de colas.

Los modelos geoestadísticos asumen que un vector $Y = (Y_1, ..., Y_n)^\top$ sigue una distribución multivariada, de modo tal que la función de densidad conjunta define una dependencia entre las variables aleatorias correspondientes. De ese modo, se puede definir un modelo estadístico a partir de dicha función de densidad. Esta tesis la dependencia espacial es incorporada en el modelo a través de cópulas gaussianas.

3.1. Definición del modelo

En esta sección se describe en detalle el modelo propuesto por Masarotto y Varin (2012). Primero se define la distribución marginal de cada v.a. Y_i de acuerdo a las características de la variable que se desea modelar. A partir de las funciones de distribución acumuladas marginales $F_{Y_i}(y_i)$, se puede usar la cópula para construir una distribución acumulada para Y. En este contexto, si se tiene una distribución de la familia exponencial $Y_i \sim FE(\mu_i, \kappa_i)$, de modo tal que $F_{Y_i}(y_i) = F_i(y_i)$ es la función de distribución acumulada de cada Y_i dónde se puede definir a $g_1(\mu_i) = x_i'\beta_x$ y $g_2(\kappa_i) = z_i'\beta_z$. Cabe resaltar que la distribución marginal puede ser otra distribución que no pertenece a la familia exponencial, pero puede ser modelada a través de su media y parámetro de escala o dispersión, por ejemplo la distribución beta (Gonzales, 2022). Luego, para construir una distribución acumulada conjunta que puede ser definida como una cópula se usa el Teorema de Slark, luego, la función de distribución acumulada conjunta puede ser definida como

$$F_Y(y_1, ..., y_n) = C(F_1(y_1), ..., F_n(y_n)).$$

Si se asume que

$$u_i = F_i(y_i),$$

donde u_i es un valor observado de $U_i \sim U(0, 1)$, entonces la cópula es una función de $(u_1, .., u_n)$., tal que:

$$F_Y(y_1, ..., y_n) = C(u_1, ..., u_n).$$

En particular, se puede asumir que C es una cópula gaussiana (ecuación 2.4), luego se tiene que:

$$F_Y(y_1, .., y_n) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)),$$

en donde Φ es una función de distribución acumulada univariada normal estándar y Φ_Σ

representa la función de distribución acumulada n-variada normal estándar con una matriz de correlación Σ. En general, esta matriz de correlación tomará en cuenta la autocorrelación entre las variables aleatorias Y_i como se define más adelante. Para ello se defienen las variables aleatorias

$$\epsilon_i = \Phi^{-1}(u_i),$$

y $\epsilon = (\epsilon_1, ..., \epsilon_n)^\top$. Entonces la función de distribución acumulada conjunta de Y puede ser redefinida por:

$$F_Y(y_1, ..., y_n) = \Phi_\Sigma(\epsilon_1, ..., \epsilon_n) = \Phi_\Sigma(\epsilon),$$

donde $\epsilon \sim N(0, \Sigma)$. Para terminar de definir la función de distribución acumulada conjunta de Y se debe definir la estructura de la matriz de correlación Σ. Gonzales (2022) aplicó la distribución beta en un modelo similar para series temporales. En esta tesis se tiene interés en datos geoestadísticos, así en particular se asume la función de correlación Matérn la cual es definida por:

$$\Sigma(y_{s_i}, y_{s_j}) = \exp(-\phi h),$$

donde h es la distancia euclideana entre los locales s_i y s_j y ϕ es un parámetro de decaimiento asociado al rango efectivo. El rango efectivo es la distancia hasta la cual dos locales aún tienen una autocorrelación espacial significativa.

Como la función de distribución acumulada conjunta de Y es construida a partir de las funciones de densidad marginales de $Y_i \sim FE(\mu_i, \kappa_i)$ y de la cópula gaussiana, implicitamente la función de distribución acumulada conjunta también depende de los coeficientes de regresión β_x y β_z.

3.1.1. Inferencia del Modelo

Los parámetros a estimar son $\Theta = (\beta, \phi)$ donde $\beta = (\beta_x, \beta_z)$. A partir de la fdp conjunta se puede definir la función de verosimilitud del modelo, la cual está dada por:

$$L(\Theta, y) = f_Y(y_1, ..., y_n),$$

$$L(\Theta, y) = f_1(y_1) f_2(y_2|y_1) ... f_n(y_n|y_{n-1} ... y_1), \tag{3.1}$$

donde $f_i(y_i) = f_{Y_i}(y_i)$.

Aunque se define inicialmente la distribución marginal de Y_i, dado que la fda conjunta de Y se construye usando la cópula, la distribución condicional de Y_i se debe hallar tomando en cuenta los errores aleatorios ϵ_i que definen la fdp conjunta de Y. Considerando el jacobiano

de la transformación de $\epsilon_i = \Phi^{-1}(F_i(y_i))$ entonces:

$$f_{y_i}(y_i|y_{i-1},...,y_1) = f_N(\epsilon_i|\epsilon_{i-1},...,\epsilon_1) \mid \frac{d\epsilon_i}{dy_i} \mid,$$

donde $f_N(\epsilon_i|.)$ es la función de densidad condicional con distribución normal.

Luego se tiene que :

$$f_{y_i}(y_i|y_{i-1},...,y_1) = f_N(\epsilon_i|\epsilon_{i-1},...,\epsilon_1)|\frac{f_{Y_i}(y_i)}{f_N(\epsilon_i)}|,$$

donde $f_N(\epsilon_i)$ es la fdp normal. Entonces reemplazando las funciones de densidad condicionales en la función de verosimilitud (ecuación 3.1) se tiene que:

$$L(\Theta; y) = \prod_{i=1}^{n} \frac{f_{Y_i}(y_i)}{f_N(\epsilon_i)} f_N(\epsilon_1,...,\epsilon_n)\,,$$

donde como se indicó en la sección anterior se asume que $\epsilon \sim N(0, \Sigma)$.

Para estimar los parámetros Θ se empleará la función de verosimilitud redefinida como:

$$L(\Theta; y) = [\prod_{t=1}^{n} f_{Y_i}(y_i)] \frac{f_N(\epsilon_1...,\epsilon_n)}{f_N(\epsilon_1)f_N(\epsilon_2)...f_N(\epsilon_n)}$$

$$L(\Theta; y) = [\prod_{t=1}^{n} f_{Y_i}(y_i)] q(\epsilon),$$

en donde de la ecuación 2.5 se puede ver que $q(\epsilon)$ representa la función de densidad de la cópula gaussiana. Luego usando la ecuación ecuación 2.6:

$$L(\Theta; y) = [\prod_{i=1}^{n} f_{Y_i}(y_i)] \frac{1}{|\Sigma^{1/2}|} \exp\{-\frac{1}{2}(\epsilon_1,...,\epsilon_n)'(\Sigma^{-1} - I)(\epsilon_1,...,\epsilon_n)\}.$$

La función de log-verosimilitud es definida por:

$$l(\Theta, y) = \frac{1}{2} log|\Sigma| + \sum_{i=1}^{n} \log f_{Y_i}(y_i) - \frac{1}{2}(\epsilon_1,...,\epsilon_n)'(\Sigma^{-1} - I)(\epsilon_1,...,\epsilon_n),$$

donde $\epsilon_i = \Phi^{-1}(F_i(y_i))$. El estimador por maxima verosimilitud (EMV) de Θ es dado por:

$$\hat{\Theta} = \text{argmax } l(\Theta; y).$$

Capítulo 4

Estudio de Simulación

El presente estudio de simulación busca evaluar el desempeño de la estimación de verosimilitud utilizando datos generados aleatoriamente de modo tal que se pueda estudiar su comportamiento con los valores reales de la simulación. Para poder llevar a cabo esta tarea se utilizaran los paquetes **geoR**, **gcmr**, **fields** y **faraway.** Asimismo, en este desempeño se analizarán para distintas distribuciones: bernoulli, poisson, gamma y beta. El detalle del estudio de esta simulación a través de sus códigos se puede ver en el anexo de esta tesis.

A continuación se detalla la implementación que se realizó para simular los datos espaciales:

1) Establecer el tamaño de la muestra n.

2) Generar n errores aleatorios ϵ_i, $i = 1, \ldots, n$, a partir de la cópula gaussiana. Es decir se generan n errores aleatorios a partir de una distribución acumulada normal, $\epsilon \sim \mathrm{N}(0, \Sigma)$ donde se usa la función de correlación exponencial (función de covarianza exponencial con $\sigma^2 = 1$). Luego, $(u_1, \ldots, u_n)^\top = \Phi_\Sigma(\epsilon)$.

3) Se simula una covariable x_i a patir de una normal estándar.

4) Se generan μ_i y κ_i, para $i = 1, \ldots, 500$ a partir de:

$$g_1(\mu_i) = \beta_{0X} + \beta_{1X} x_i,$$
$$g_2(\kappa_i) = \beta_{0Z}.$$

   En el caso de la distribución binomial $\pi_i = \mu_i$ y en el caso de la distribución poisson $\lambda_i = \mu_i$, respectivamente según esta notación general.

5) Finalmente se simulan los valores observados de la v.a. Y_i, para $i = 1, \ldots, 500$, mediante

la inversa de la fda de una distribución expecífica (binomial, poisson, gamma o beta), tal que, $y_i = F_{Y_i}^{-1}(u_i)$, para $Y = (Y_1, \ldots, Y_n)^\top$.

Luego el objetivo es estimar los posibles parámetros $\Theta = (\beta_{0x}, \beta_{1x}, \beta_{0z}, \phi)$.

4.1. Detalle de la generación de errores aleatorios espaciales

Para la generación de los errores aleatorios espaciales se va a utilizar el paquete **geoR** propuesto por Peter Diggle (2007) y dentro de esta la función *grf()*. Esta función nos permite elegir como se va a estructurar la correlación. Así, para la generación de los errores aleatorios se sigue el siguiente proceso. Sea $Z = (Z_1, ..., Z_n)$ la realización de números aleatorios gaussianos estándar, es decir $Z_i \sim N(0, 1)$. Luego aplicamos una transformación lineal:

$$S = AZ,$$

en donde S es un proceso gaussiano $\{S(x) : x \in \mathbb{R}^2\}$, A es una matriz cuyo valor es $AA' = \Sigma$. Es decir, se va a utilizar una trasformación de Cholesky para su simulación. Por otro lado, la función nos permite definir la estructura de correlación de los datos a simular, definida de forma general de la siguiente forma:

$$C(h) = \sigma^2 \rho(h),$$

donde σ^2 es la varianza marginal y $\rho(h)$ la autocorrelación entre dos variables cuya distancia entre ellas es h. Para el estudio de esta simulación se está tomando en cuenta una función de covarianza exponencial. De ese modo, nuestra función de correlación para analizar será la siguiente:

$$\Sigma = \rho(h) = \exp(-\phi h),$$

donde el parámetro $\phi = 2/r$ es un parámetro de decaimiento asociado al llamado rango efectivo (r), es decir la distancia hasta la cual la autocorrelación espacial es significativa.Finalmente, como datos adicionales, los datos se generan en un espacio [0,1] (Figura 4.1) y los parámetros adicionales como la varianza (σ^2) y el parámetro r. Se tomaran como 1 y 0.15 respectivamente.

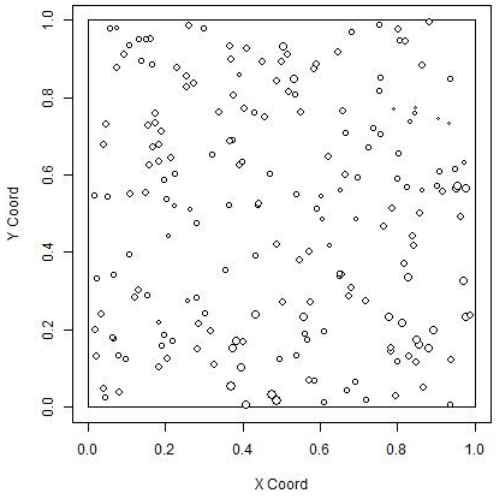

Figura 4.1: Puntos Generados

Del mismo modo, la Figura 4.2 muestra el variograma para los datos simulados, en donde se puede analizar la influencia de un punto sobre otros basándose en la ubicación espacial.

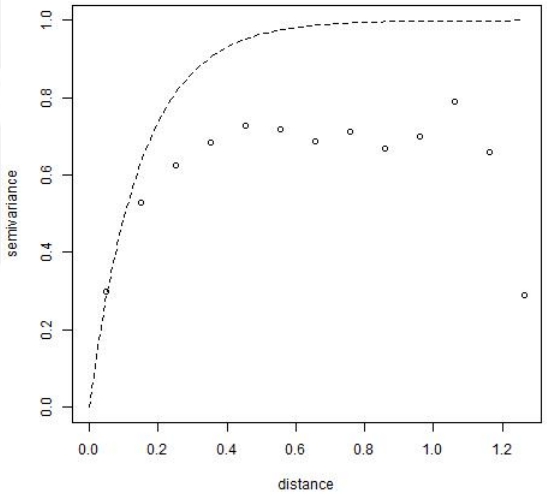

Figura 4.2: Variograma

4.2. Resultados de las simulaciones

En la presente sección se evaluarán como se comportan las distribuciones en una prueba de simulación. Para dicho objetivo se utilizará el paquete **gcmr** en donde se estiman las regresiones con cópulas a través de máximo de verosimilitud. Esta seccíon cuenta con dos estudios. En primer lugar, se prueban los resultados de la data generada con su respectiva simulación así como sus intervalos de confianza relacionados. Se busca estimar los parámetros de las siguientes regresiones: bernoulli, poisson, gamma y beta.

El cuadro 4.1 muestra los resultados obtenidos de la simulación para las diferentes ditribuciones explicadas en el desarrollo del modelo. En primera instancia, se observa que los valores originales son muy parecidos a los valores tras la estimación así como los valores originales se encuentran dentro del intervalo de confianza.

Cuadro 4.1: Resultados de simulación

Distribución	Parámetro	Valor Original	Estimación	IC
binomial	β_{0x}	1	1.061***	(0.895 - 1.227)
	β_{1x}	1	0.960***	(0.906 - 1.0145)
	r	0.15	0.144***	(0.117-0.171)
poisson	β_{0x}	3	2.989***	(2.980 - 2.999)
	β_{1x}	2.5	2.505***	(2.500 - 2.509)
	r	0.15	0.158***	(0.135 - 0.181)
gamma	β_{0x}	2	2.018***	(1.858 - 2.177)
	β_{1x}	1	1.005***	(0.923 - 1.087)
	β_{z0}	4	3.996***	(3.530 - 4.462)
	r	.15	0.167 ***	(0.086 - 0.2468)
beta	β_{0x}	-2.4	-2.410***	(-2.514 - -2.305)
	β_{1x}	-0.14	-0.142***	(-0.151 - -0.134)
	β_{z0}	6	6.006 ***	(5.541 - 6.471)
	r	0.15	0.167 ***	(0.086 - 0.249)

Sin embargo, este resultado no termina siendo suficiente para poder analizar la bondad de ajuste de los modelos geoestadísticos usando cópulas gaussianas. Es por ello que se realiza en segunda instancia una simulación en donde se va a usar el mismo tamaño de la muestra planteada en un inicio; sin embargo, este experimento será realizado N veces.

En el cuadro 4.2 se observa que a medida que la cantidad de repeticiones se incrementan, el sesgo de estos valores se va a ir reduciendo y a través de ello los valores estimados se van a acercando más a los valores originales empleados para la simulación. Por ejemplo si se compara los resultados obtenidos en el sesgo de N=100 contra el sesgo en N=1000 esta

última tiene los valores más cercanos a cero tanto para los coeficientes de regresión como para el resultado del rango efectivo asociado a los errores aleatorios dentro del proceso gaussiano.

Cuadro 4.2: Resultados de réplicas

			N=100	N=500	N=1000
	Parámetro	Valor Original	sesgo	sesgo	sesgo
binomial	β_{0x}	1	-0.0117	0.13	0.000
	β_{1x}	1	0.0019	0.00	-0.000
	r	0.15	-0.0043	-0.03	-0.03
poisson	β_{0x}	1	0.26	0.13	0.01
	β_{1x}	1	-0.03	0.00	-0.00
	r	0.15	0.03	-0.03	-0.03
gamma	β_{0x}	1	0.26	0.13	0.01
	β_{1x}	1	-0.03	0.00	-0.00
	β_{0z}	4	-0.02	-0.003	0.003
	r	0.15	0.056	-0.002	-0.0000
beta	β_{0x}	3	-0.003	-0.0007	0.01
	β_{1x}	2.5	0.002	0.000	-0.00
	β_{0z}	6	0.01	0.002	0.001
	r	0.15	-0.0057	-0.0078	-0.0003

Capítulo 5

Aplicaciones

En la presente sección se busca realizar evaluaciones empíricas de los modelos presentados en la sección anterior. Así entonces, se ha optado por utilizar dos ejemplos aplicativos. En primer lugar se está realizando una estimación bernoulli (caso particular de una binomial) a través de la estimación de las probabilidades de encontrar petróleo en el dominio espacial de interés. Asimismo, se usa una distribución gamma para estimar la contaminación ambiental en otro dominio espacial de interés.

5.1. Aplicación 1

El conjunto de datos a utilizar se refiere a perforaciones exitosas o no exitosas en la búsqueda de yacimientos pretroleros. Estos puntos de ubicación se encuentran en Delaware, Nuevo México. Asímismo, el tratamiento de estos datos ha sido realizado por Han y Oliveira (2018).

Como se puede observar en la Figura 5.1 existe una relación entre los porcentaje de éxito en la búsqueda de yacimientos petrolíficos. Dentro del los cuadrantes correspondientes a *Northing* entre 17 y 24 así como en los níveles 17 a 30 en *easting* de 17 a 30 es donde más probabilidad existen. No es casualidad que a medida que los valores se van alejando de esta zona, la probabilidad se va reduciendo. Es decir, se observa que hay un vínculo con la dependencia espacial entre los puntos.

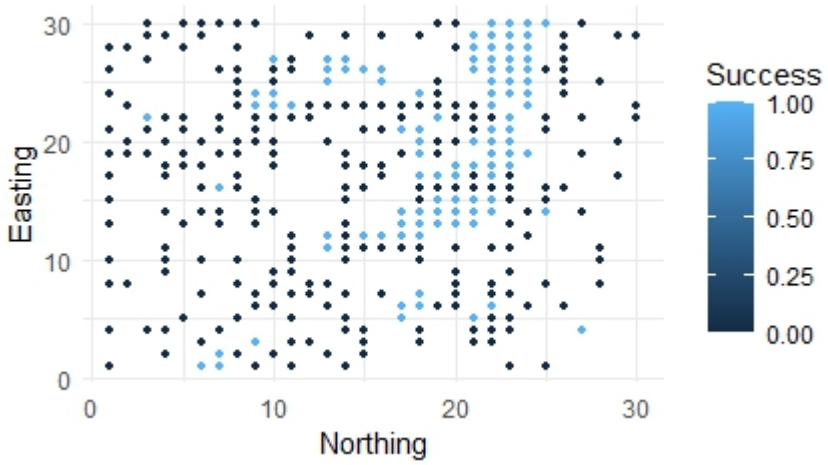

Figura 5.1: Probabilidad Éxito Yacimientos

Asimismo, si se analiza el variograma de los datos se puede apreciar en la figura 5.2, que el rango es pequeño, aproximadamente 2.

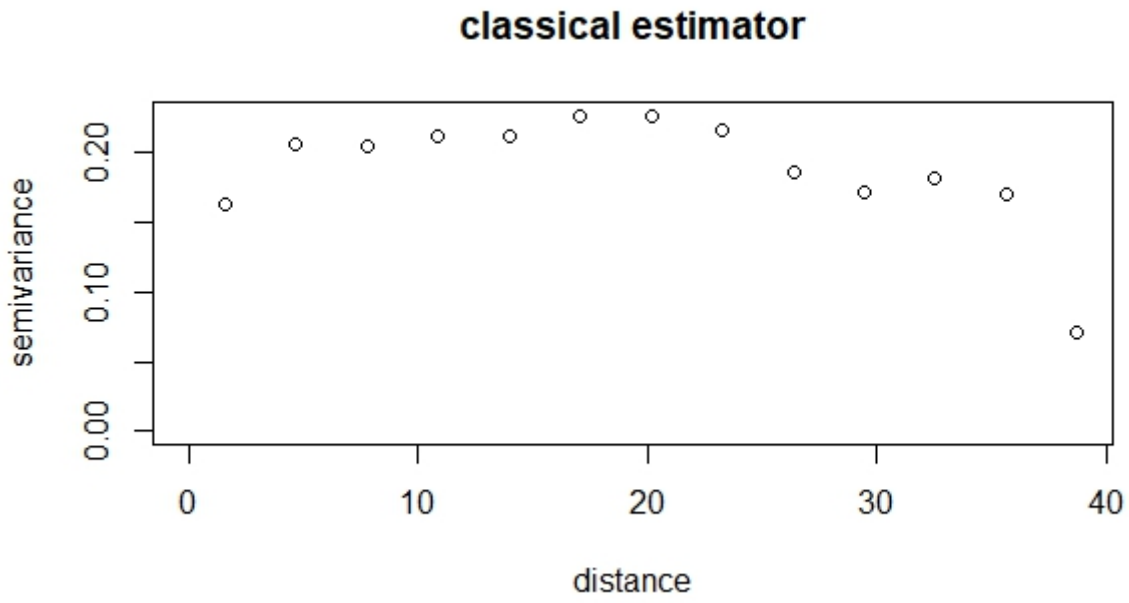

Figura 5.2: Variograma Yacimiento

Ahora bien para realizar la estimación, se asume que $Y_i = 1$ si la perforación petrolífera es exitosa, y por otro lado $Y_i = 0$ si la perforación petrolífera no es exitosa. Entonces se asume que marginalmente $Y_i \overset{ind}{\sim} ber(\pi_i)$, donde π_i representa la probabilidad de que la perforación

petrolífera sea exitosa. Luego se define:

$$g_1(\pi_i) = logit(\pi_i) = \beta_{0X} + \beta_{1X}x_{1i} + \beta_{2X}x_{2i} + \beta_{3X}x_{3i},$$

donde las covariables x_{1i}, x_{2i} y x_{3i} están representadas por las covariables northing, northing^2 y easting, respectivamente. En efecto, como se revisó en la parte descriptiva, hay una mayor importancia en la unbicación hacia el norte. Por lo que se utilizó como covariable adicional el cuadrado de la variable northing. Para incorporar la estructura espacial en el modelo se procede a aplicar el modelo geoestadístico usando una cópula gaussiana. Los resultados se presentan en el cuadro 5.1.

Cuadro 5.1: Resultados de simulación

	Estimación	Error Estándar	Z-value	$Pr(>\|z\|)$
β_{0X}(Intercepto)	-6.605394	1.650820	-4.001	0.000***
β_{1X} $(Northing)$	0.645540	0.188687	3.421	0.000***
β_{2X} $(Northing^2)$	-0.0188371	0.005635	-3.260	0.001***
β_{3X}(Easting)	0.03205	0.33005	0.971	0.331522

En los resultados del cuadro 5.2 se puede apreciar que las dos variablers *Northing* son relevantes para el modelo. Sin embargo, se puede observar que existe un valor negativo en el caso cuadrático. Esto resulta muy interesante cuando se contrasta con los datos de la Figura 5.1. En ella se puede apreciar que efectivamente en los extremos del norte probabilidad de éxito es menor. Específicamente como $\exp(-0{,}018) = 0{,}982161$ por cada incremento en una unidad de northing al cuadrado, la chance de que la perforación petrolífera sea exitosa se reduce en 1.78 %. Esta característica la estamos rescatando del modelo pues el valor de *northing* al cuadrado permite explicar este comportamiento. Por otro lado la variable easting indica que por cada incremento de unidad de la variable easting, la chance de que una perforación petrolífera sea exitosa se incrementa en 3.25 %.

Asímismo, en el cuadro 5.2 se puede apreciar el valor del rango efectivo r dentro de la cópula. Se puede apreciar que el valor del rango efectivo r es relevante para el modelo, es decir el parámetro de la cópula. En particular, se puede decir que el resultado de una perforación petrolífera en un local, depende espacialmente de forma significativa del resultado en otro local hasta una distancia de 1.6732 unidades.

Cuadro 5.2: Resultados de simulación

	Estimación	Error Estándar	valor-z	$Pr(>\|z\|)$
r	1.6732	0.3105	5.389	0.000***

Finalmente, se observa la capacidad de estimación de la probabilidad de estimación. De ese modo, combinando los valores estimados de la probabilidad de perforación petrolífera exitosa con la existencia de yacimientos pretrolíferos tras la exploración, los resultados se muestran en la figura 5.3. Y usando un punto de corte de 0.3, para identificar si una perforación petrolífera es exitosa o no, se estima correctamente el 78.07 %.

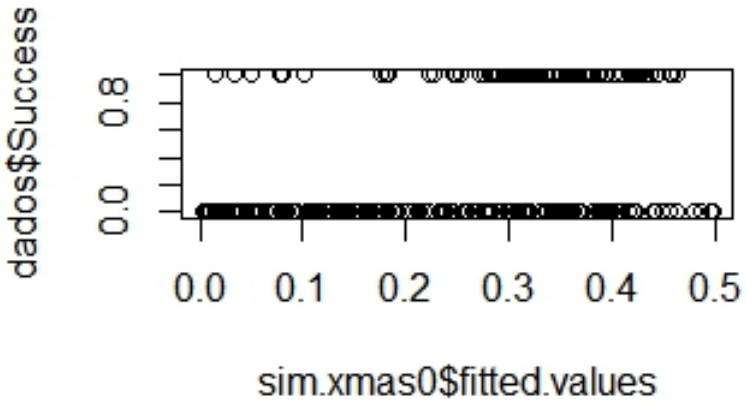

Figura 5.3: Predicción del modelo

5.2. Aplicación 2

En esta aplicación se usaron datos recolectados por UK AIR - Air (Information Resource) sobre contaminación ambiental en el Reino Unido (`https://uk-air.defra.gov.uk`). La variable de interés es la concentración de material particulado $PM_{2,5}$ en el reino unido en el 2020 en estaciones urbanas y rurales. Esta variable nos permite medir el grado de contaminación ambiental

El objetivo es poder estimar el material particulado en el reino unido. A partir de nuestro análisis exploratorio, el $PM_{2,5}$ se encuentra en un intervalo de 0 a 12.5. Es una variable que solo toma valores positivos y su distribución es ligeramente asimétrica.

En la figura 5.4 se puede identificar en primera instancia si se encuentra en una zona rural o en una zona urbana. Del mismo modo, se ve el nivel de contaminación en el aire.

Esto es que cercano a Londres los niveles de contaminación se va a incrementando. Mientras más se aleja hacia zonas fuera de las capitales, la contaminación se reduce. Esto nos da la indicación que la distancia respecto a un punto, como Londres, juega un papel importancia en el cálculo del nivel de contaminación.

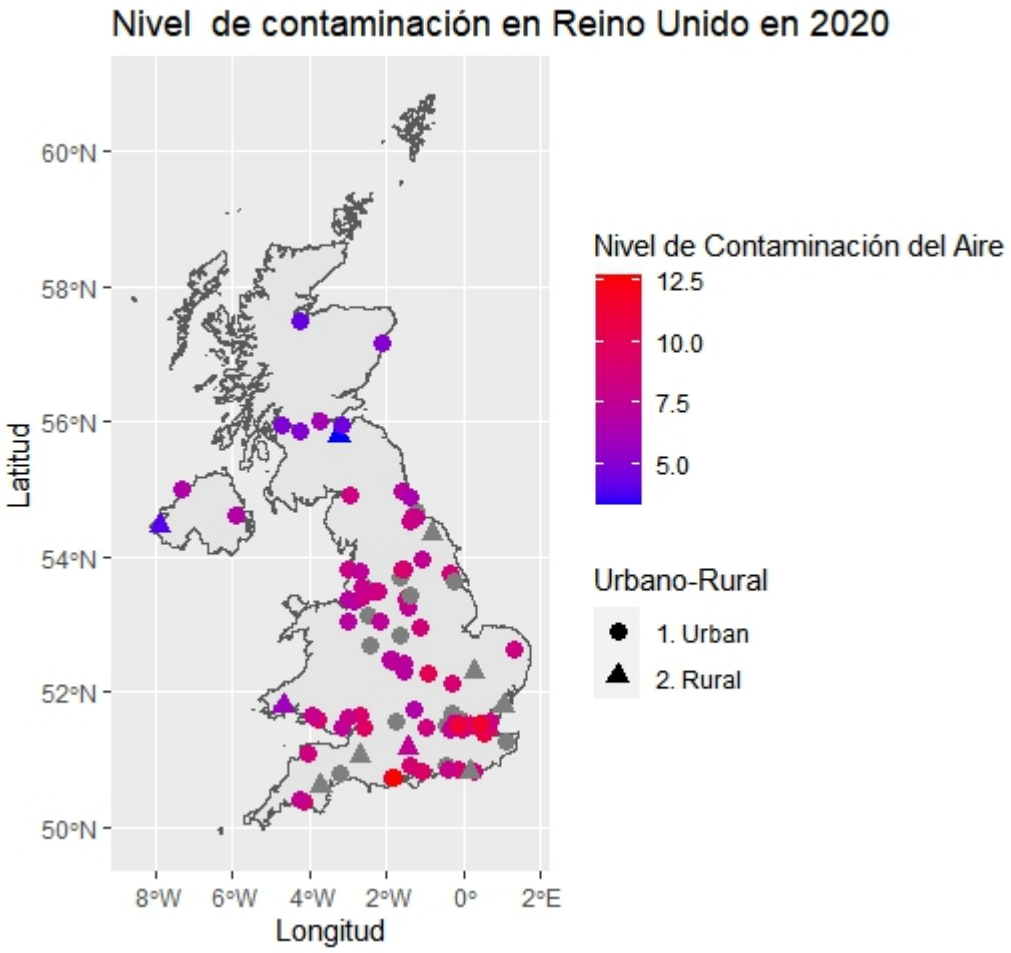

Figura 5.4: Mapa de contaminación del aire según zonas rurales y urbanas.

Se define la v.a. Y_i, que representa el valor de $PM_{2,5}$ en una estación s_i, $i = 1, 2, \ldots, 80$. Como se indicó la variable es ligeramente asimétrica. (figura 5.5). Se prueba que la variable no sigue una distribución normal a través de una prueba de Shapiro-Wilk. Ahora bien, es necesario encontrar una distribución que nos permita modelar como la variable dependiente una variable que tome valores positivos. Es así que se va a emplear una distribución marginal gamma.

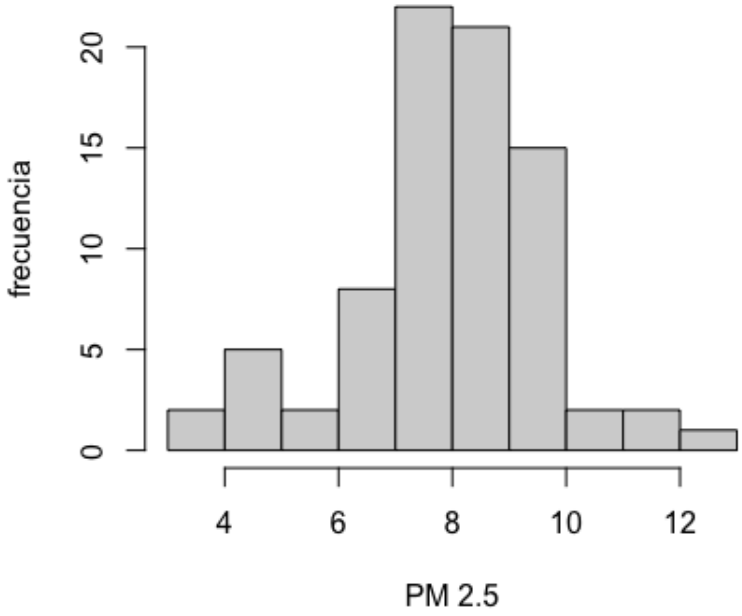

Figura 5.5: Histograma de material particulado 2.5.

Luego se asume que $Y_i \overset{ind}{\sim} gamma(\mu_i, \kappa_i)$, donde μ_i representa la media de $PM_{2,5}$ y κ_i es un parámetro de dispersión. Luego se define:

$$g_1(\mu_i) = \beta_{0X} + \beta_{1X} x_{1i} + \beta_{2X} x_{2i} + \beta_{3X} x_{3i},$$

$$g_2(\kappa_i) = \beta_{0z},$$

donde las covariables x_{1i}, x_{2i} y x_{3i} están representadas por las covariables latitud, rural (si o no) y longitud, respectivamente. Y las funciones de enlace g_1 y g_2 son la función inversa y logarítmica, respectivamente.

Se puede ver en el variograma (figura 5.6) la evidencia de dependencia espacial entre los valores de $PM_{2,5}$. Se podría decir que según este resultado exploratorio el rango efectivo es alrededor de cuatro unidades.

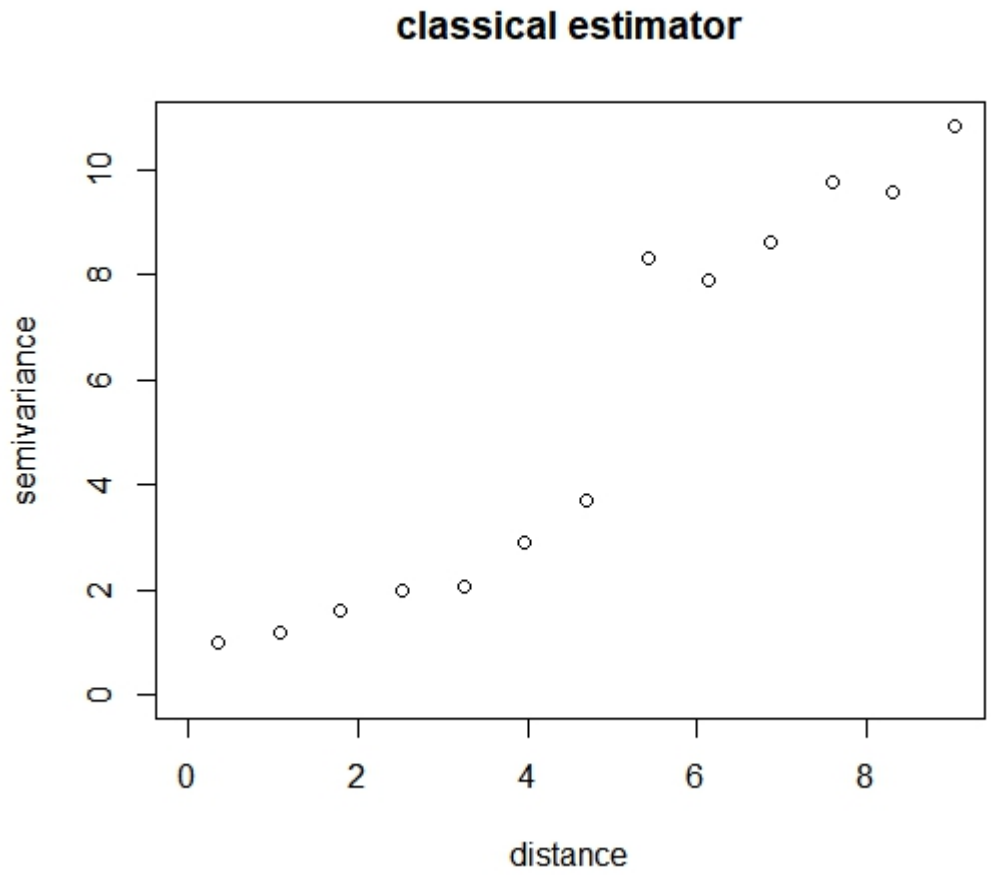

Figura 5.6: Variograma de las distancias

Para incorporar la estructura espacial en el modelo se procede a aplicar el modelo geoestadístico usando una cópula gaussiana. Los resultados se presentan en el cuadro 5.3.

Cuadro 5.3: Resultados de aplicación 2

	Estimación	Error Estándar	Z-value	$Pr(>\|z\|)$
β_{0X} $(Intercepto)$	-0.4108	0.000	-89.73	0.000***
β_{1X} $(Latitud)$	0.0098	0.000	113.64	0.000***
β_{2X} $(DummyRural)$	0.0194	0.000	27.20	0.000***
β_{3X} $(Longitud)$	-0.0057	0.000	-48.44	0.000***
β_{0z}	7.742	0.000	481.81	0.000***
r	2.70917	0.00131	2068	0000*

En cuanto al caso de la significancia de los coeficientes de regresión se ve que todos estos son signficativos dentro del modelo propuesto para estos datos. También se puede apreciar que la presencia en zonas rurales ha afectado la contaminación referente a los resultados

obtenidos en las zonas urbanas. Específicamente, si una zona es rural, el valor de $PM_{2,5}$ se reduce en 0.019. Del mismo modelo, se ve que por cada una en que se incrementa la latitud, el $PM_{2,5}$ se reduce en 0.0098 unidades. También se concluye que por cada una en que se incrementa la longitud, el $PM_{2,5}$ se incrementa en 0.005 unidades. El rango efectivo es alrededor de 2.7 unidades, por lo tanto la variable $PM_{2,5}$ en un local depende de forma significativa de los valores de esta variable hasta un radio de 2.7 unidades.

En la figura 5.7 Se puede apreciar en el caso de los valores estimadas en la regresión que existe una buen ajuste del modelo, pues las estimaciones de $PM_{2,5}$ usandoel modelo propuesto son similares a los valores reales de $PM_{2,5}$.

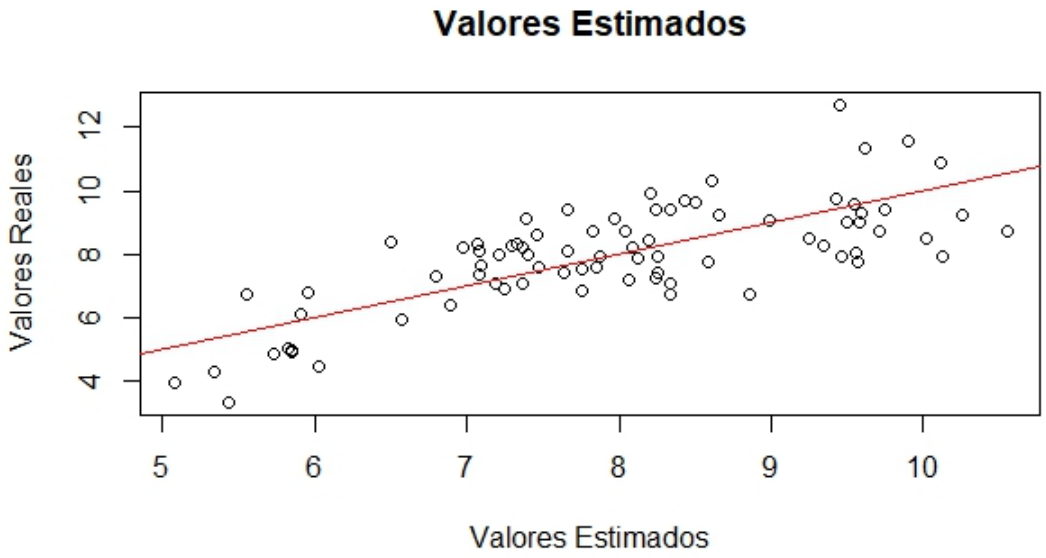

Figura 5.7: Valores estimados de $PM_{2,5}$ usando el modelo propuesto versus valores observados de $PM_{2,5}$.

Capítulo 6

Conclusiones

Se evidencia que existe una dependencia entre puntos dentro de un mismo espacio. Esta dependencia puede llegar a influir en los valores que se quieren analizar o en las hipótesis sobre los datos que se quiere extraer. De ese modo, entre distintas aproximaciones que ofrece la literatura se ha observado el uso de cópulas gaussianas que puedan ser incluidas para modelar la dependencia espacial en la regresión.

Las cópulas permite lidiar con la no linealidad o normalidad tomando en cuenta la dependencia entre las variables de respuesta. Asi, se ha utilizado el modelo de Masaroto et al. (2012). Se ha estudiado a través de simulaciones sus resultados para cuatro tipo de distribuciones: binomial, poisson, gamma y beta. Se obtienen estimaciones de los parámetros consistentes y poco sesgados así como valores estimados caen dentro de los intervalos de confianza. De ese modo, una vez establecida que la consistenta de estos modelos es válida, se optó por utilizar dos evaluaciones empiricas. En primer lugar se buscó una prueba para una distribución bernoulli y en segundo lugar se utilizó para una distribución gamma.

El primer ejemplo consiste en encontrar la existencia de petróleo al momento de realizar una excavación. Los resultados obtenidos nos sugieren que es importante tomar en cuenta la latitud así como la longitud del modelo. Es importante resaltar que el modelo propuesto toma en cuenta la autocorrelación espacial, de esta forma que puede estimar el valor del rango efectivo, y así saber si se realiza una excavación sobre qué radio se debería realizar excavaciones.

En el segundo caso se esudia la contaminación ambiental en reino unido a través de data obtenida de ciudades tanto urbanas como rurales. Para ese cálculo se utiliza una variable positiva que es el material particulado 2.5. Entonces se optó por utilizar una distribución gamma que responde a esas características en la variable dependiente. Por otro lado, dentro del modelo también se buscó identificar si existen alguna diferenciación entre ciudades

urbanas o ciudades rurales. Se encontró que la ubicación espacial termina afectando a las ciudades rurales.

6.1. Sugerencia para investigaciones futuras

En base a lo estudiado en el presente documento, se puede obtener una agenda de investigación relacionada al uso de modelos espaciales para el estudio climatológico en el Perú así como otros temas aplicados como competencia en el sector minorista entre otras cosas. En cuanto al aspecto metodológico, se considera que el modelo debe de probarse con más distribuciones y utilizar distintos tipos de cópulas de modo tal que el desempeño del modelo al enfrentarse a problemas puntuales en los datos sea mucho más óptimo.

Apéndice A

Código R

Se muestran los códigos utilizados en la simulaciones y las Aplicaciones.

A.1. Simulación

A.1.1. Simulación Distribución Benroulli

```
set.seed(666) # Semilla
beta_0x= 1
beta_1x= 1

n= 100 #Total de Observaciones
m= 100 #Total de Simulaciones

b0_100=rep(0,m)
b1_100=rep(0,m)
tau_100=rep(0,m)

control=1

for ( i in 1:m){
  sim1 <- grf(n, cov.pars = c(1, .15))
  loc = sim1$coords
  D <- spDists(cbind(loc[,1], loc[,2]))
  w = sim1$data
  u <- pnorm(w)
  x1=rnorm(n,0,1)
  logitmut = beta_0x +beta_1x*x1
  mut =ilogit(logitmut)
```

```
  Ntrials = sample(1:10, size=n, replace=TRUE)
  z2 <-  qbinom(u, size = Ntrials, mut)
  sim.xmas1 <- gcmr( cbind(z2, Ntrials-z2)  ~ x1 | 1 ,
                     marginal = binomial.marg,
                      cormat = matern.cormat(D))
  b0_100[i]= sim.xmas1$estimate[1]
  b1_100[i]= sim.xmas1$estimate[2]
  tau_100[i]=sim.xmas1$estimate[3]
}

tabla_betas_n_100 <- cbind(b0_100,b1_100,tau_100)
write.csv(tabla_betas_n_100,"tabla_betas_n_100.csv")

jpeg("Simulaciones con N=100 datos.jpeg")
par(mfrow=c(1,3))
boxplot(b0_100,col = "lightgray",main="b0 N=100")
boxplot(b1_100,col = "bisque",main="b1 N=100")
boxplot(tau_100,col = "blue",main="tau N=100")
dev.off()

sesgo_beta_b0_100=beta_0x-mean(b0_100)
sesgo_beta_b1_100=beta_1x-mean(b1_100)
sesgo_beta_tau_100=0.15-mean(tau_100)


# N= 300 Observaciones
n= 300 #Total de Observaciones

b0_300=rep(0,m)
b1_300=rep(0,m)
tau_300=rep(0,m)

control=1

for ( i in 1:m){

  sim1 <- grf(n, cov.pars = c(1, .15))
  loc = sim1$coords
  D <- spDists(cbind(loc[,1], loc[,2]))
  w = sim1$data
  u <- pnorm(w)
  x1=rnorm(n,0,1)
```

```
  logitmut = beta_0x +beta_1x*x1
  mut =ilogit(logitmut)
  Ntrials = sample(1:10, size=n, replace=TRUE)
  z2 <-  qbinom(u, size = Ntrials, mut)
  sim.xmas1 <- gcmr( cbind(z2, Ntrials-z2)  ~ x1 | 1 ,
  marginal = binomial.marg,
                      cormat = matern.cormat(D))
  b0_300[i]= sim.xmas1$estimate[1]
  b1_300[i]= sim.xmas1$estimate[2]
  tau_300[i]=sim.xmas1$estimate[3]
}

tabla_betas_n_300 <- cbind(b0_300,b1_300,tau_300)
write.csv(tabla_betas_n_300,"tabla_betas_n_300.csv")

jpeg("Simulaciones con N=300 datos.jpeg")
par(mfrow=c(1,3))
boxplot(b0_300,col = "lightgray",main="b0 N=300")
boxplot(b1_300,col = "bisque",main="b1 N=300")
boxplot(tau_300,col = "blue",main="tau N=300")
dev.off()

sesgo_beta_b0_300=beta_0x-mean(b0_300)
sesgo_beta_b1_300=beta_1x-mean(b1_300)
sesgo_beta_tau_300=0.15-mean(tau_300)

# N= 500 Observaciones

n= 500 #Total de Observaciones
m= 100 #Total de Simulaciones

# Valores iniciales
beta_0x= 1
beta_1x= 1



b0_500=rep(0,m)
b1_500=rep(0,m)
tau_500=rep(0,m)

control=1
```

```
for ( i in 1:m){

  sim1 <- grf(n, cov.pars = c(1, .15))
  loc = sim1$coords
  D <- spDists(cbind(loc[,1], loc[,2]))
  w = sim1$data
  u <- pnorm(w)
  x1=rnorm(n,0,1)
  logitmut = beta_0x +beta_1x*x1
  mut =ilogit(logitmut)
  Ntrials = sample(1:10, size=n, replace=TRUE)
  z2 <-  qbinom(u, size = Ntrials, mut)
  sim.xmas1 <- gcmr( cbind(z2, Ntrials-z2)  ~ x1 | 1 ,
  marginal = binomial.marg,
                     cormat = matern.cormat(D))
  b0_500[i]= sim.xmas1$estimate[1]
  b1_500[i]= sim.xmas1$estimate[2]
  tau_500[i]=sim.xmas1$estimate[3]
  control=control+1
  print(control)
}

tabla_betas_n_500 <- cbind(b0_500,b1_500,tau_500)
write.csv(tabla_betas_n_500,"tabla_betas_n_500.csv")

jpeg("Simulaciones con N=500 datos.jpeg")
par(mfrow=c(1,3))
boxplot(b0_500,col = "lightgray",main="b0 N=500")
boxplot(b1_500,col = "bisque",main="b1 N=500")
boxplot(tau_500,col = "blue",main="tau N=500")
dev.off()

sesgo_beta_b0_500=beta_0x-mean(b0_500)
sesgo_beta_b1_500=beta_1x-mean(b1_500)
sesgo_beta_tau_500=0.15-mean(tau_500)
```

A.2. Modelo Gamma

```
 install.packages("qtl2")
install.packages("dplyr")
```

```
library(qtl2)
library(dplyr)
filtered_data <- read.csv(file = "filtered_data.csv")

all_sites <- filtered_data %>%
  group_by(site) %>%
  summarize(mean = mean(pm2.5, na.rm = TRUE),
  latitude = first(latitude), longitude = first(longitude),
  site_type = first(site_type))
head(all_sites, 5)

all_rural_data <- all_sites %>%
  filter(site_type == "Rural Background")
head(all_rural_data, 3)

transform_site_type <- function(site_type) {
  result <- c()
  for (i in (1:length(site_type))) {
    if ((site_type[i] == "Urban Background") |
    (site_type[i] == "Urban Industrial") |
    (site_type[i] == "Urban Traffic")) {
      result <- c(result, "1. Urban")
    } else {
      result <- c(result, "2. Rural")
    }
  }
  result
}

all_sites <- all_sites %>%
  mutate(new_site_type = transform_site_type(site_type))
head(all_sites)

library(rgeoboundaries)

# Download the boundaries
uk_boundary <- geoboundaries(country = "GBR")

# Plot the map
ggplot(data = uk_boundary) +
  geom_sf() +
  geom_point(data = all_sites, aes(x = longitude, y = latitude,
```

```
  shape = new_site_type, color = mean), size = 3) +
  scale_color_gradient(name = "Nivel de Contaminaci n del Aire",
  low = "blue", high = "red") +
  scale_shape_discrete(name = "Urbano-Rural") +
  ggtitle("Nivel promedio de contaminaci n en Reino Unido en 2020") +
  labs(x = "Longitud", y = "Latitude")

# variogram
dados=all_sites
library(geoR)
datos2=as.geodata(cbind(dados$longitude,dados$latitude,(dados$mean)))

bin1 <- variog(datos2)

plot(bin1, main = "classical estimator")

library( gcmr )

library("sp")

D <- spDists(cbind(dados$longitude, dados$latitude))
#image.plot(D)

names(dados)

indNA = which(is.na(dados$mean)==TRUE)
dados= dados[-indNA,]

shapiro.test(dados$mean)
# como en nuestro caso es menor que 0.05, rechazamos normalidad.

sim.xmas0 <- gcmr(  mean ~  latitude  + new_site_type + longitude| 1,
                    data = dados, marginal =Gamma.marg(link = "inverse"),
                    cormat = matern.cormat(D))

summary(sim.xmas0)


cbind(sim.xmas0$fitted.values, dados$mean)

plot(cbind(sim.xmas0$fitted.values, dados$mean))
abline(0,1,col="red")
```

Bibliografía

Bai, Y., Kang, J. y Song, P. (2014). Efficient pairwise composite likelihood estimation for spatial-clustered data, *Biometrics* **70**(3): 661–70.

Cressie, N. (1993). *Statistics for Spatial Data*, John Wiley Sons.

Cribari-Neto, F. (2006). Improved point and interval estimation for a beta regression model, *Computational Statistics & Data Analysis* **51**(15): 960–981.

Gonzales, A. R. C. (2022). *Regresión beta usando cópulas gaussianas para analizar series de tiempo*, Master's thesis, Pontificia Universidad Católica del Perú, Perú.

Han, Z. y Oliveira, V. D. (2018). On the correlation structure of Gaussian copula models for geostatistical count data, *Australian and New Zealand Journal of Statistics* **58**(1): 47–69.

Joe, H. (2014). *Dependence Modelling with Copulas*, Chapman Hall.

Kazianka, H. y Pilz, J. (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates, *Stochastic Environmental Research and Risk Asessment* **24**: 661–673.

Krupskii, P. y Genton, M. G. (2018). Linear factor copula models and their properties, *Biometrics* **45**(4): 861–878.

Kurowicka, D. y Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence*, Wiley.

Masarotto, G. y Varin, C. (2012). Gaussian copula marginal regression, *Electronic Journal of Statistics* **6**: 1517 – 1549.
URL: *https://doi.org/10.1214/12-EJS721*

Masarotto, G. y Varin, C. (2017). Gaussian copula regression in R, *Journal of Statistical Software* **77**(8): 1–26.
URL: *https://www.jstatsoft.org/index.php/jss/article/view/v077i08*

Nelsen, R. (2007). *An Introduction to Copulas*, Springer.

Peter Diggle, P. R. (2007). *Model-based Geostatistics*, Springer.

Pitt, M., Chan, D. y Kohn, R. (2006). Efficient bayesian inference for Gaussian copula regression models, *Biometrika* **93**(3): 537–554.

Song, P., Li, M. y Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas, *Biometrics* **65**(1): 60–68.